Searching for species in haloarchaea

R. Thane Papke*†, Olga Zhaxybayeva†, Edward J. Feil‡, Katrin Sommerfeld†, Denise Muise†, and W. Ford Doolittle†§

*Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT 06269-3125; †Department of Biochemistry and Molecular Biology, Dalhousie University, 5850 College Street, Halifax, NS, Canada B3H 1X5; and ‡Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, United Kingdom

Contributed by W. Ford Doolittle, July 10, 2007 (sent for review June 1, 2007)

Prokaryotic (bacterial and archaeal) species definitions and the biological concepts that underpin them entail clustering (cohesion) among individuals, in terms of genome content and gene sequence similarity. Homologous recombination can maintain gene sequence similarity within, while permitting divergence between, clusters and is thus the basis for recent efforts to apply the Biological Species Concept in prokaryote systematics and ecology. In this study, we examine isolates of the haloarchaeal genus *Halorubrum* from two adjacent ponds of different salinities at a Spanish saltern and a natural saline lake in Algeria by using multilocus sequence analysis. We show that, although clusters can be defined by concatenation of multiple marker sequences, barriers to exchange between them are leaky. We suggest that no nonarbitrary way to circumscribe ''species'' is likely to emerge for this group, or by extension, to apply generally across prokaryotes. Arbitrary criteria might have limited practical use, but still must be agreed upon by the community.

Halorubrum | homologous recombination | multilocus sequence analysis | species definition

Genomics and metagenomics are breathing new life into old debates about prokaryotic species. At issue are (*i*) whether microbes naturally form cohesive genotypic or phenotypic clusters, (*ii*) how to recognize such clusters, and (*iii*) when they deserve the status of "species." Linked to these questions about natural pattern and species definition is another: what ecological, genetic, and evolutionary processes are responsible for clustering, if and when it does occur? Two types of species concept address this problem of process.

In ecotype models (1), cohesion is achieved by periodic selection between clones within ecologically defined, primarily asexual populations (''ecotypes''). When advantageous new mutant alleles sweep to fixation, the rest of the genome in which they first arose hitchhikes to high frequency, because rates of homologous recombination (HR) are too low to disrupt this linkage. Genetic cohesion within ecotypes thus entails periodic purging of diversity at all loci. Divergence between ecotypes, on the other hand, is a consequence of their genetically determined ecological distinctness, which might arise from just one or a few genetic differences that prevent a genome that sweeps to fixation in one ecotypical niche from invading another, even when sympatric. While maintaining internal cohesion, ecotypes evolve and diverge.

The alternative, Ernst Mayr's Biological Species Concept (BSC), was first applied to bacteria in 1991 (2) and engenders much current excitement in bacterial population genetics. The BSC assumes that within-population recombination is frequent: it is genes, not whole genomes, that achieve fixation as populations evolve. HR is indeed much more common among bacteria than we had thought just a few years ago, as demonstrated through whole genome comparisons and metagenomic community studies (e.g., ref. 3) but most extensively, convincingly, and quantifiably through multilocus sequence analysis (MLSA) (4). In MLSA, 5–10 housekeeping genes are sequenced from scores to hundreds of strains, and the extent to which recombination must be invoked to explain the spectrum of allelic profiles (''sequence types'' or STs) is assessed. HR occurs at widely varying rates but is evident among almost all taxa, and for many is the major cause of sequence divergence.

Whether or not HR can be the basis of a realistic BSC-like species model that ensures divergence between clusters (speciation) as well as cohesion within clusters depends importantly on the existence, nature, and effectiveness of barriers preventing HR between their genes. Ecological distinctness alone is not sufficient, because it only precludes recombination within or near the loci responsible for it. Nor is it necessary, if other barriers are effective. Candidates for such other barriers include simple physical separation (allopatry), host specificity of DNA exchange systems (for instance, plasmids, phages, and DNA uptake, modification, and restriction systems), and stringent requirements of the recombinational machinery for sequence similarity between donor and recipient.

It is appealing to base speciation models on this last phenomenon. Given genetic mechanisms already understood, HR rates should fall off rapidly as sequences diverge, as has been observed in several experimental systems (5). Fraser *et al.* (6) show by computer modeling that, if HR varies appreciably between members of a population, species-like cohesion coupled with between-species divergence might result, even in sympatric situations. However, these conditions are sufficiently special that speciation should more often occur as a consequence of allopatry, niche specialization or some equivalent hindrance to DNA transfer (for instance, limited host range of plasmids and phages, as mentioned above). Also, Fraser *et al.* (6) caution that possible selection is not taken into account in their model; nor, we note, is the fact that different genes diverge at different rates during speciation. Thus, recombination may still be frequent at some loci while having effectively ceased at others.

Only real data can tell us whether, in Nature, recombining bacteria do form populations sufficiently cohesive and bounded that we might want to call them species under the BSC. On this topic, there are just a few case studies using MLSA. Hanage *et al.* (7), examining species of the genus *Neisseria*, showed that single-locus trees (for each of seven housekeeping genes) are incongruent and fail to reproduce recognized species clusters, but concatenated MLSA data do group together almost all strains assigned to them by traditional methods. Such species may be "real," but are "fuzzy." Either alleles have frequently exchanged between them or, less probably, there has not been time since their separation as diverging populations for all ancestral polymorphisms to have gone to fixation or extinction. A similar result (genuinely incongruent single-gene phylogenies

Author contributions: R.T.P. designed research; R.T.P., O.Z., K.S., and D.M. performed research; R.T.P., O.Z., E.J.F., and W.F.D. analyzed data; and R.T.P., O.Z., E.J.F., and W.F.D. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: BSC, Biological Species Concept; HR, homologous recombination; ST, sequence type; MLSA, multilocus sequence analysis; SLV, single locus variant.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AM777162–AM777382 and AM777425–AM777776).

§To whom correspondence should be addressed. E-mail: ford@dal.ca.

This article contains supporting information online at www.pnas.org/cgi/content/full/0706358104/DC1.

but unique well resolved trees for concatenated data) is obtained for *Streptococcus pneumoniae* and relatives (8).

For more distantly related taxa (sister genera, for instance), one might expect less fuzziness, and indeed this was observed by Wertz *et al.* (9) for representatives of six named genera. Here, each of five housekeeping genes reproduced the same monophyletic species clusters, as if there were no recombination between them. Relationships between species were nevertheless different for the different genes, a result that could reflect recombination between species at an earlier stage (before the last common ancestors of strains currently comprising them) or (less likely, we think) rapid radiative speciation of a polymorphic common ancestral population.

There are so far just three model systems for studying HR and its implications for Archaea. First, Whitaker *et al.* (10), looking at six loci among 60 *Sulfolobus islandicus* isolates from Kamchatka, concluded that frequent HR "prevents periodic selection from purging diversity and provides a fundamental cohesive mechanism within this and perhaps other archaeal species." Second, Banfield and colleagues (3) infer that a population of the acidophilic archaeon *Ferroplasma acidarmanus* at an acid mine drainage site in California "is undergoing frequent genetic recombination, resulting in a mosaic genome pool that is shaped by selection." Third, in a pilot study of a saltern in Santa Pola, Spain, Papke *et al.* (11) found that cooccurring strains of *Halorubrum* sp. with identical 16S rRNA genes experienced HR so frequent that the population approached linkage equilibrium (panmixis).

Haloarchaea such as *Halorubrum* are excellent models for field and laboratory study of the "species question" as it pertains to Archaea. They are physiologically diverse, have dynamic genomes, and show "island biogeography" because of the patchy distribution of hypersaline waters. High-intensity UV light at such sites induces expression of the recombinational machinery (12) and mating occurs naturally, possibly via intercellular cytoplasmic bridges (13). Growth conditions, although extreme, are easy to replicate, tractable surrogate genetics systems exist (14), and recombination readily occurs between introduced and chromosomal markers. Here, we extend the pilot *Halorubrum* study (11), describing a more extensive sampling and MLSA analysis of population structure at two Santa Pola sites and a third site that is separated from them by 250 km. Multiple ribotypes within the *Halorubrum* cluster were included. We find both within-population cohesion, defining three "phylogroups," and between-population exchange, eroding such cohesion.

Results

***Halorubrum* Isolates, Genes, and Alleles.** We examined five loci from 153 strains assigned to *Halorubrum* on the basis of 16S rRNA gene sequence and cultivated from three hypersaline sites: two adjacent ponds of 22% and 36% salinity from a saltern near Alicante, Spain, and an inland Algerian hypersaline lake ≈250 km away (22% saline). In an effort to recover the most similar inhabitants from these three sites, all were isolated on plates at 25% salinity. The most frequently recovered 16S rRNA allele was identical to *Halorubrum* sp. strain Aus-1.

Supporting information (SI) Tables 1–6 summarize most data discussed here. Sequenced fragments ranged from 305 to 507 nt, the number of alleles per locus ranged from 12 to 46, and the number of polymorphic sites (positions at which variation was observed) ranged from 24 to 114. The 16S rRNA locus was most conserved in number of polymorphic sites, number of alleles, and heterozygosity (as defined in SI Table 2). The bacteriorhodopsin-encoding *bop* had the most polymorphic sites, and *radA* had the highest number of alleles. For the protein coding loci, *dN*/*dS* ratios indicated that each was under purifying selection (SI Table 1). Many alleles were observed more than once, and the most frequent were observed in 27–87 isolates. Distributions of allele

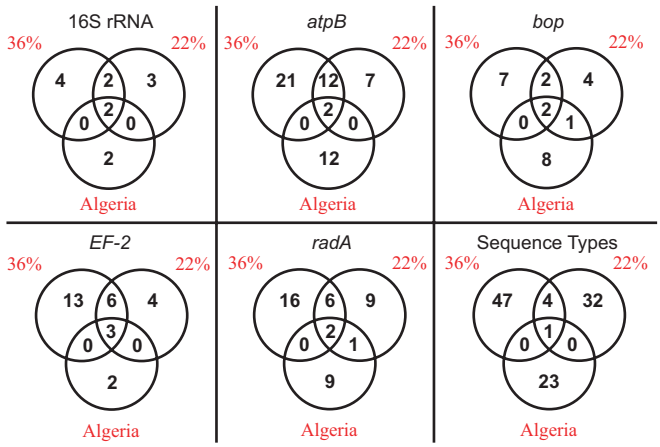

Fig. 1. Venn diagrams of allele and ST distributions. Individual circles represent 36% and 22% salinity Spanish saltern ponds and the Algerian site. Values inside the circles reflect number of alleles or STs in each set.

frequencies conform to nearly neutral expectations, with a few very common alleles and an excess of rare alleles observed once. Determination of the probability of drawing different alleles at random from the sample ("heterozygosity") showed that the 16S rRNA gene was most conserved and *radA*, a *recA* homolog, most diverse (SI Table 2). Interestingly, the three "housekeeping" protein-coding loci were more diverse than *bop*, which is patchily distributed among haloarchaea and subject to lateral gene transfer (15). The 153 strains define 104 STs: the most frequent occurred only 12 times, and the vast majority were unique.

Distribution of Alleles and STs Between Sampling Sites. The two Spanish saltern ponds differ in salinity (22% and 36%) but are only meters apart. The geographically isolated Algerian site is of a very similar salinity to one of the Spanish sites (22%). In total, 28 alleles are shared between the Spanish sites but absent from the Algerian site, whereas only two alleles are found only in the Spanish 22% site and the Algerian site, and no alleles were common only to the Spanish 36% site and the Algerian site (Fig. 1). In terms of STs, four were noted exclusively in the two Spanish sites, but none was noted in the Algerian site and only one of the Spanish sites. Thus, the two Spanish sites are more similar to each other in terms of allele frequency than either is to the Algerian site: geographical proximity may explain more of the between-site similarities than does adaptation to salinity.

Phylogenetic Analyses: Defining Phylogroups. The phylogenetic tree in Fig. 2 is based on concatenating all five loci, and illustrates the broad *Halorubrum* diversity captured here. Several clades are supported by significant bootstrap values (≥70%), and in particular three large clusters are revealed, here designated phylogroups A, B, and C. Together, these clades incorporate the majority (86%; 89 of 104) of the STs in the sample. The robustness of these clades is underlined by the fact they differ by an average of 3.2–5.7% nucleotide divergence, whereas the average nucleotide divergence within each clade is ≤1.2% (SI Table 3). Using a 99% cutoff value for the average nucleotide identity (ANI) analysis (16), which provides a uniform criterion for circumscribing clusters, largely preserved these three phylogroups, although it excluded STs at the base of clusters (e.g., ST052, ST053, ST113, ST082, and ST097) and split phylogroup C in two by isolating sequence types ST131, ST132, and ST151 (SI Table 7). Application of >94% ANI and >97% 16S rRNA identity cutoffs lumped together all STs except those in phylogroup Z. Although 99% ANI has been recently suggested by Konstantinidis and Tiedje (16) as more comparable to species

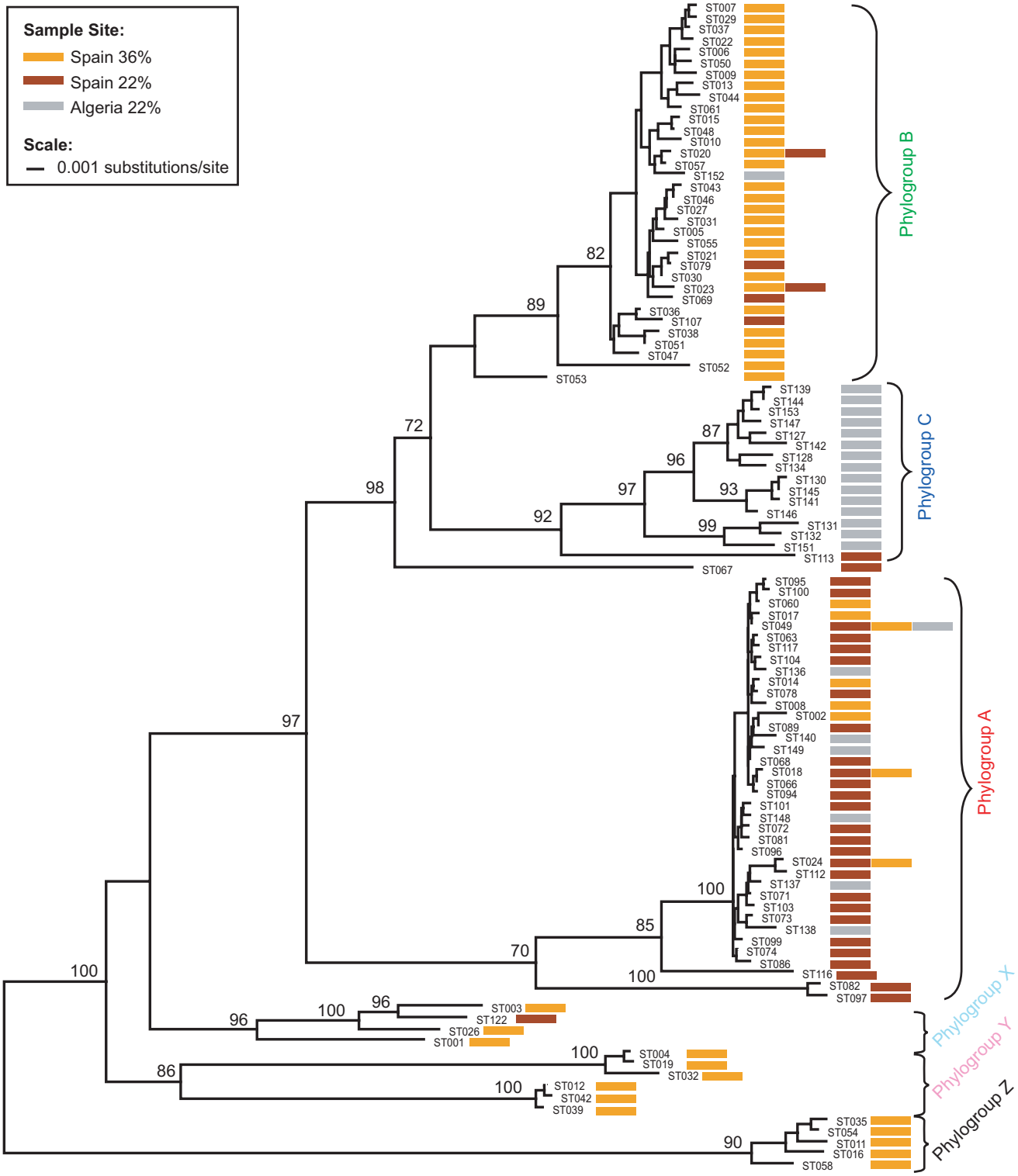

Fig. 2. Concatenated gene phylogeny of *Halorubrum* STs. Color bars indicate where strains representing STs were obtained. STs with multiple color bars have individual strains cultivated from different sites. Phylogroups are defined as relatively tight clades on this tree. Phylogroup Z was chosen as outgroup based on knowledge of the 16S rRNA gene phylogeny of the haloarchaea. Only bootstrap support values >70% are shown.

definitions in the rest of biology, the lower values correspond to widely used criteria for defining prokaryotic species (1).

The Venn diagrams in Fig. 1 give some understanding of the genetic data and their distribution among the sample sites, but examining the same information in the light of phylogroups provides a more robust picture. There is a good mapping of phylogroups to sample sites, but there is some sharing of STs in all major phylogroups between sites (Fig. 2). Phylogroup A is primarily (78%) found in the lower salinity Spanish pond, for instance, but eight (≈12%) of its STs are found at higher salt and seven are Algerian, whereas two STs are found in both Spanish locations, and one is found at all three sites. Phylogroups B and

C, although the closest pair phylogenetically, are the most distinct ecologically when salinity and location are both considered.

Recombination Between Phylogroups Assessed by Phylogenetic Incongruence. In several instances, incongruent phylogenetic topologies were detected when trees for individual loci were compared with the concatenated gene tree (SI Fig. 4 and SI Table 4). For example, with the exception of EF-2, all protein-coding loci from ST116 form a strong relationship with typical members of phylogroup A, whereas the "incongruent" locus forms a strong relationship (100% bootstrap support) with ST001, an ST from phylogroup X. Comparison of the ST116 EF-2 allele to a "typical" phylogroup A EF-2 allele (e.g., ST002) revealed changes at 29 nucleotides, a ≈6% divergence. SI Table 4 summarizes the notable instances of such incongruent alleles, which involved nearly 8% of the STs from phylogroups A, B, and C, and ranged from 19 to 63 nucleotide changes (4–17% divergence).

The 16S rRNA gene also displayed evidence of phylogenetic incongruence (SI Fig. 4*A*). For instance, ST001, a member of phylogroup X, had a 16S rRNA allele identical to several STs from phylogroup C. Indeed, none of the phylogroups identified in the concatenated phylogeny, except for the distantly related phylogroup Z (used as outgroup), is monophyletic in the 16S rRNA tree (SI Fig. 4*A*), as if this gene were more frequently involved in recombination than are the protein-coding loci.

It is important to realize that disagreements between individual trees genuinely reflect different histories. Even when few nucleotide substitutions separate alleles, it is unlikely that we are being misled by homoplasies (recurrent identical patterns of mutation), because substitutions are rare overall. Thus, it is not the case that there is a single true phylogeny for these genes, which is captured by the concatenated sequence but obscured by noise in the individual analyses. Moreover, when STs fall at the base of the phylogroups as we have delimited them, this is as often due to recombination with some even more distant lineage affecting a single locus as it is to mutational divergence at all loci (see SI Fig. 5 *B–E* and SI Table 4).

There is also evidence for between-phylogroup recombination before the divergence of the within-phylogroup lineages available in our sampling. Although most alleles (the above-mentioned 8% excluded) from each of the loci recovered the same phylogroups, individual genes do not all support the same relationships between these phylogroups. For each of the loci, we examined the relationships among the different phylogroups by using "likelihood mapping" (17), and found conflict between them (Fig. 3). For instance, a strong relationship between phylogroups B and C was obtained for the EF-2 and *radA* loci, whereas the *bop* locus clearly favored a relationship between phylogroups A and C. The *atpB* locus, on the other hand, revealed a complex evolutionary history; a majority of the time, there was strong support for the B–C relationship, but a significant fraction supported a robust A–C relationship. Such incongruent phylogenies most probably reflect between-phylogroup recombination earlier in the histories of the phylogroups. A less likely explanation is unsorted polymorphisms in the population ancestral to all three phylogroups.

Splits decomposition analyses can also be used to assess the support of data for a strictly bifurcating tree (18). If a data set contains conflicting phylogenetic signals (supporting contradicting splits), a network will result. We calculated a neighbor-net (see SI Fig. 6) for the concatenated data set. Although phylogroups A, B, and C can be generally outlined, some sequences appear clearly outside the groups (e.g., ST113, ST082, and ST097).

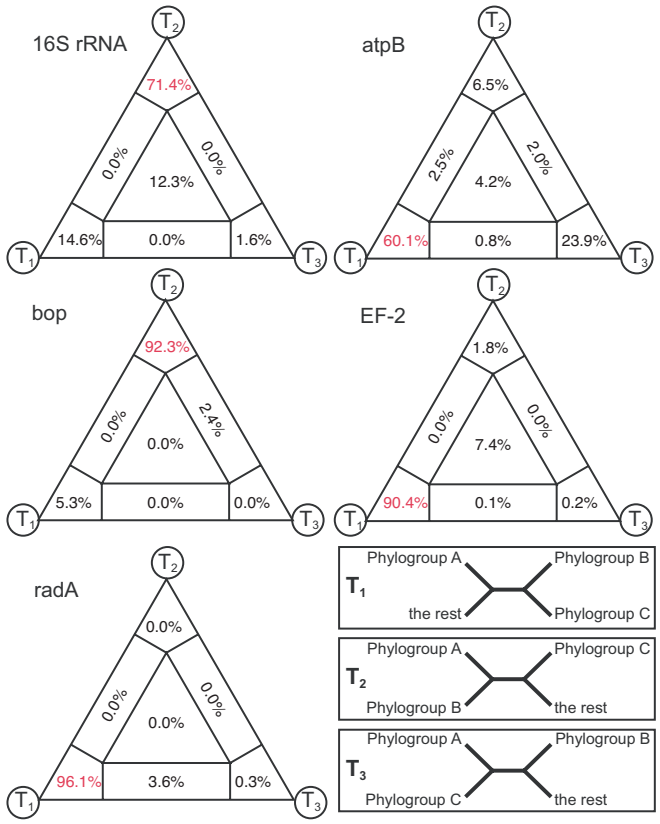

Fig. 3. Maximum likelihood mapping analysis of phylogenetic signal within individual loci. STs (Fig. 2) were divided into four groups corresponding to phylogroups A, B, and C and the rest of the taxa, and three possible phylogenetic relationships (T_1, T_2, or T_3) among these groups were evaluated by using all possible quartets. Support for each relationship and each locus is summarized in barycentric coordinates. In this coordinate system, topologies are placed at the vertices of an equilateral triangle whose area is divided into regions. Values inside each region show the percentage of quartets that fall into the region (and the quartet location is determined by the likelihood of supporting each topology). Values in the regions closest to a vertex show percentage of quartets that strongly support the tree topology at the vertex. Values along edges indicate the percentage of quartets unable to discriminate between the two possible topologies located at the adjacent vertices. Values in the middle indicate the percentage of quartets that did not discriminate between the three topologies. When more than one topology is strongly supported (e.g., 16S rRNA), we interpret this as evidence for highly supported conflicting relationships within a locus, as opposed to noise, which is reflected in the values in the middle region. Notably, different loci support different relationships among the tested groups.

The Relative Contributions of Recombination and Mutation to Diversification. Inconsistencies due to recent allelic replacements by HR manifest in two contrasting ways: diversification or homogenization. In a single pair of strains, strikingly different alleles may be observed within one gene, while all other genes remain identical. In this case, HR acts as a diversifying force. Alternatively, a single identical allele may be shared between two otherwise highly divergent strains; here, recombination is a cohesive force.

The diversification of MLSA genes may proceed either by homologous recombination or by point mutation. Considering very closely related strains, where only one of the MLSA genes has changed, greatly simplifies the problem of multiple hits obscuring the footprints of these two processes. We used a clustering algorithm, eBURST (19), to demarcate lineages and identify pairs of strains differing at only a single locus (single locus variants; SLVs). In eBURST, each unique allele is assigned

a number and STs are defined by allelic profiles (SI Table 8). Using a threshold of three out of five identical alleles to demarcate groups, eBURST parsed the STs into nine "clonal complexes" and eight "singletons." Complexes 1, 2, and 3 (SI Fig. 7) encompassed 79% of the total strains and were roughly identical in composition to phylogroups A, B, and C, respectively (SI Table 7). eBURST also recreates parsimonious short-term patterns of descent based on a model of radial diversification from clonal founders (SI Fig. 7). All pairs of strains differing at only a single locus (SLVs) were identified by using eBURST. Variant alleles were treated as putative point mutations if differing by a single base, and as recombinational replacement if differing by at least two bases (on average approximately five per event). This is a simplified version of a previous approach for gauging relative contributions of these two processes.

Of 42 SLV pairs examined, 15 showed a single base pair difference, and 27 revealed multiple base pair differences; the per allele recombination/mutation ratio is ≈2:1 in favor of recombination. The 27 alleles assigned to recombination accounted for a total of 128 nucleotide changes; the per-site recombination/mutation ratio is thus ≈9:1. Of these 27 alleles, 25 corresponded to changes in eight or fewer nucleotides. The two exceptions were a *bop* allele exchange between STs 132 and 151, which resulted in 12 nucleotide changes, and an EF-2 allele exchange (STs 074 and 116) that resulted in 28 nucleotide changes (SI Table 8). A comparison of the distribution of nucleotide differences between alleles differing in SLVs and all pairwise comparisons of alleles in the protein coding genes reveals significantly fewer nucleotide changes in SLVs. Two factors are responsible: the enrichment of recent point mutations when SLVs are considered (accounting for the excess of single nucleotide changes) and the fact that homologous recombination is more likely between closely related strains belonging to the same phylotype. The only exceptions are the two diverse replacements discussed above; thus, gene flow is highly structured, and the total sample comprises more than one subpopulation.

In structured populations with nonrandom mating, knowledge of one allele predicts a second unrelated allele (linkage disequilibrium). By using the index of association (I_A) (20) statistic to determine the randomness of allele distribution, we tested several groupings that might correspond to natural population substructures (SI Table 6). Linkage equilibrium within such subpopulations could be shown when subpopulations were defined either as phylogroups, eBURST complexes, or STs with identical 16S rRNA alleles. However, the I_A values for 16S rRNA-defined populations were not very close to zero and probably reflect their patchy distribution with respect to the concatenated gene alignment phylogeny (Fig. 2). Pairwise combinations of phylogroups and eBURST complexes (e.g., A+B), were in disequilibrium, which emphasized their incomplete mixing.

Intragenic Recombination Assessed from Aligned Sequences. HR events need not respect gene boundaries, and recombination can also be assessed for aligned gene and genomes sequences by using a variety of algorithms. For each locus, the PHI test (21) found statistically significant evidence for intragenic recombination (SI Table 5). Visual inspection of alignments revealed that each protein-coding locus had at least one and often multiple alleles that originated in different phylogroups (see SI Fig. 5). For instance, the *radA* locus from ST131 appears to be a mosaic of phylogroup C and A, and the EF-2 locus from ST004 appears to be a mosaic of phylogroups A, B, and C, and possibly other unknown species.

Discussion

Distribution of Phylogroups. The different-salinity ponds at the Spanish site have been intensely studied by Rodríguez-Valera and colleagues (22, 23) and contain fundamentally distinct bacterial and eukaryal communities; we expected a similar result for *Halorubrum*. With concatenated gene sequences defining phylogroups, a statistically significant localization of groups to sample site indeed emerges: phylogroup A members are far more likely to be found in the 22% salinity Spanish pond, phylogroup B members prefer the 36% Spanish site, and phylogroup C is almost exclusively Algerian.

Nevertheless, migration does occur between sites, because phylogroups A and B include STs found at all sample sites, and the Algerian location contains representatives of all three phylogroups. Moreover, phylogenies based on individual genes show that migration of phylogroup C must have occurred, even though STs in phylogroup C appear very largely confined to Algeria. For instance, it is clear that the *bop* allele of ST113, the only non-Algerian member of phylogroup C, is Algerian in origin: it is intermingled with Algerian *bop* loci in its phylogenetic tree (see SI Fig. 4) and differs in at least 19 nucleotide positions from any other *bop* collected in Spain. ST113's EF-2 and *radA* alleles are also likely to have come from a phylogroup C member or close relative, whereas its *atpB* locus associates it with Spanish isolates (see SI Fig. 4). Furthermore, the *radA* locus appears to have been involved in an intragenic recombination with some lineage outside phylogroups A and B (SI Fig. 5*E*). Additional evidence for the migration of phylogroup C alleles comes from the observation that members of the Spanish phylogroup X either have identical 16S rRNA sequences to members of phylogroup C or differ from them by but a single nucleotide (SI Fig. 4*A*).

Data such as these harbor both biogeographic and phylogenetic signals. For instance, we did not detect isolates in Spain with a full complement of phylogroup C alleles (grouping only with other C alleles in trees). It is likely that there is a leaky barrier to dispersal and migration is slow with respect to evolutionary change; everything is not everywhere at this level of resolution. Geographic isolation will almost certainly play a more important role in the diversification of *Halorubrum* over distances of >250 km.

Recombination, Diversification and Adaptation. The mosaic nature of ST113 highlights a conclusion of this study: HR, rather than mutation, drives sequence diversification in *Halorubrum.* In addition, it is possible that, in haloarchaea, as in some bacteria, modulation of the rate at which recombination can occur with DNA from relatively unrelated lineages is part of the evolutionary dynamic. In several bacteria, the mismatch repair system limits recombination between divergent sequences. When it is inactivated (as in "mutator strains"), more distantly related sequences can be more readily assimilated by HR. Such more distant alleles (already screened by stabilizing selection in the donor genome) are more likely to offer selectively significant functional differences (24). ST067 may have such a mutator phenotype. From Fig. 2, ST067 has high bootstrap support for being related to phylogroups B and C, but falls into neither. Inspection of individual phylogenies reveals a highly varied phylogenetic relationship for each locus (SI Table 4 and Fig. 4). Additionally, two protein-coding loci in this ST are implicated in intragenic recombination events (e.g., *atpB* and *bop*) and two (EF-2 and *radA*) have multiple (six and five, respectively) unique nucleotide changes, for which recombination with unknown donors is the most likely explanation (SI Fig. 5).

When HR introduces an advantageous allele from outside a population or assembles a particularly advantageous combination of alleles from within, it can also drive adaptation. In the extreme, the genome that first acquires such an allele or combination will sweep to fixation, as in Cohan's periodic selection model (25). The preponderance of certain STs within phylogroup A may indicate such events in progress, or completed and now suffering erosion through HR. Even when HR is so frequent

as to frustrate such genome-level purging of diversity, it will still be the case that selection reduces diversity at the targeted loci. In this study, we noticed that the *bop* locus in phylogroups A and B has little diversity compared with other loci (SI Fig. 5*C*). In both instances, *bop* had only two synonymous polymorphic sites, whereas, for instance, in phylogroups A and B, *radA* has 11 and 9 polymorphic sites, respectively. Furthermore, *bop* was second only to the functionally very conservative 16S rRNA gene in terms of lowest heterozygosity and number of alleles (SI Table 1). Many haloarchaea, including strains of *Halorubrum*, do not produce bacteriorhodopsin, and it was recently shown by phylogenetic methods that *bop* gene loss and replacement are common (15). Recurrent periodic sweeps driven by *bop* genes introduced via LGT into bop$^-$ populations adapting to anaerobiosis would explain low diversity at this locus compared with our other markers, which are essential as genes and not subject to the same evolutionary dynamics.

Are Phylogroups Species? Several recent MLSA or otherwise multigenic studies of bacterial populations take the high bootstrap values of trees obtained with concatenated gene data as evidence for the existence of true species (3, 8, 9). The rationale seems to be that the greater resolution of concatenates justifies treating the incongruence of trees for the individual genes whose sequences were strung together as the equivalent of phylogenetic noise. Such attempts to reify "species" seem to us misguided. First, bootstrap values can increase with increased data even when that data includes genuinely incongruent signals (26, 27). Second, we already know that much of the gene data are, in fact, genuinely incongruent; disagreement is not "noise," as this term is commonly understood. Third, as long as populations are incompletely mixed, alleles will be differently distributed among them. The more loci are examined collectively, the more reliably we will be able to assign individuals to subpopulations. However, this does not mean that these subpopulations have the level of cohesion we expect of species, indeed that they may not be almost completely mixed. Similar, and similarly contentious, would be any attempt to unambiguously circumscribe human "races" from the observation that, with a sufficient number of SNPs, one can identify continents of origin of many individual humans (28).

Surely, microbial populations are incompletely mixed; it would defy common sense to claim that 250 km or a 60% difference in salinity represent no barrier whatsoever to gene exchange. Even when cohabiting the same cubic centimeter of saltern water and using the same substrates, genetic exchange can be reduced between cells that have different sensitivities to infection by the phages or invasion by the plasmids that might be the primary agents of gene exchange. However, there is no reason for such barriers not to vary continuously from completely ineffective to completely effective in preventing exchange, and no reason then not to expect degrees of cohesion of microbial assemblages also to vary continuously, differing from group to group for purely contingent and often temporary reasons.

Whether we should call *Halorubrum* phylogroups A, B, and C "species" or simply subpopulations cannot be decided until we agree on some uniform measures and standards of cohesion. Most of the debates over "species definitions" address what might be the most practically useful measure of within-cluster similarity, not what degree of clustering (within-group similarity together with between group dissimilarity, the latter requiring that there are no missing intermediates) might actually exist. Nor do such debates tackle the issue of how uniformly across the microbial world this degree of clustering needs to be observed before we can say that the category species is real and that each and every individual bacterial or archaeal cell can be properly said to belong to one and only one species. (An alternative would be to say that some belong to species and some do not.)

As Hanage *et al.* (8) recently remarked of the claim that clusters we can call bacterial species exist, "In fact, there are almost no data that address this assertion, which in essence is a statement of belief. A more agnostic view is to ask whether populations of similar bacteria do invariably (or usually) form discrete well-resolved genotypic clusters that merit the status of species and to consider which methods should be used to address this issue." We suggest that concatenation does not address the issue satisfactorily, because it will inevitably produce clusters as long as there is any degree of geographic or ecological structuring of bacterial populations. What we expect in terms of discreteness of such clusters before we will call them species remains to be negotiated. Until we have agreed on what we are looking for, we cannot tell whether we have found it.

Methods

Strain isolation, cultivation, PCR amplification, and sequencing were performed as described in ref. 11 and in detail in *SI Methods*. Phylogenetic analyses, likelihood mapping, and SplitsTree methods, as well as assessment of recombination, are also described in *SI Methods*.

We thank Karima Kharroub, Arantxa Lopez-Lopez, and Francisco Rodriguez-Valera for help with sampling, and Canadian Institutes of Health Research (CIHR) (Grant MOP-4467) and Genome Atlantic for support. O.Z. is a CIHR and honorary Killam Postdoctoral Fellow.

1. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J (2005) *Nat Rev Microbiol* 3:733–739.
2. Dykhuizen DE, Green L (1991) *J Bacteriol* 173:7257–7268.
3. Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF (2007) *Proc Natl Acad Sci USA* 104:1883–1888.
4. Feil EJ, Spratt BG (2001) *Annu Rev Microbiol* 55:561–590.
5. Majewski J, Cohan FM (1999) *Genetics* 153:1525–1533.
6. Fraser C, Hanage WP, Spratt BG (2007) *Science* 315:476–480.
7. Hanage WP, Fraser C, Spratt BG (2005) *BMC Biol* 3:6.
8. Hanage WP, Fraser C, Spratt BG (2006) *Philos Trans R Soc London B* 361:1917–1927.
9. Wertz JE, Goldstone C, Gordon DM, Riley MA (2003) *J Evol Biol* 16:1236–1248.
10. Whitaker RJ, Grogan DW, Taylor JW (2005) *Mol Biol Evol* 22:2354–2361.
11. Papke RT, Koenig JE, Rodriguez-Valera F, Doolittle WF (2004) *Science* 306:1928–1929.
12. McCready S, Muller JA, Boubriak I, Berquist BR, Ng WL, Dassarma S (2005) *Saline Systems* 1:3.
13. Rosenshine I, Tchelet R, Mevarech M (1989) *Science* 245:1387–1389.
14. Soppa J (2006) *Microbiology* 152:585–590.
15. Sharma A, Walsh D, Bapteste E, Rodriguez-Valera F, Doolittle W, Papke R (2007) *BMC Evol Biol* 7:79.
16. Konstantinidis KT, Tiedje JM (2005) *Proc Natl Acad Sci USA* 102:2567–2572.
17. Strimmer K, von Haeseler A (1997) *Proc Natl Acad Sci USA* 94:6815–6819.
18. Huson DH, Bryant D (2006) *Mol Biol Evol* 23:254–267.
19. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG (2004) *J Bacteriol* 186:1518–1530.
20. Smith JM, Smith NH, O'Rourke M, Spratt BG (1993) *Proc Natl Acad Sci USA* 90:4384–4388.
21. Bruen TC, Philippe H, Bryant D (2006) *Genetics* 172:2665–2681.
22. Benlloch S, Lopez-Lopez A, Casamayor EO, Ovreas L, Goddard V, Daae FL, Smerdon G, Massana R, Joint I, Thingstad F, *et al.* (2002) *Environ Microbiol* 4:349–360.
23. Casamayor EO, Massana R, Benlloch S, Ovreas L, Diez B, Goddard VJ, Gasol JM, Joint I, Rodriguez-Valera F, Pedros-Alio C (2002) *Environ Microbiol* 4:338–348.
24. Townsend JP, Nielsen KM, Fisher DS, Hartl DL (2003) *Genetics* 164:13–21.
25. Cohan FM (2001) *Syst Biol* 50:513–524.
26. Gadagkar SR, Rosenberg MS, Kumar S (2005) *J Exp Zool B* 304:64–74.
27. Jermiin LS, Poladian L, Charleston MA (2005) *Science* 310:1910–1911.
28. Allocco DJ, Song Q, Gibbons GH, Ramoni MF, Kohane IS (2007) *BMC Genomics* 8:68.